On the benefits for model regularization of a Variational formulation of GTM

Iván Olier and Alfredo Vellido

Abstract— **Generative Topographic Mapping (GTM) is a manifold learning model for the simultaneous visualization and clustering of multivariate data. It was originally formulated as a constrained mixture of distributions, for which the adaptive parameters were determined by Maximum Likelihood (ML), using the Expectation-Maximization (EM) algorithm. In this formulation, GTM is prone to data overfitting unless a regularization mechanism is included. The theoretical principles of Variational GTM, an approximate method that provides a full Bayesian treatment to a Gaussian Process (GP)-based variation of the GTM, were recently introduced as alternative way to control data overfitting. In this paper we assess in some detail the generalization capabilities of Variational GTM and compare them with those of alternative regularization approaches in terms of test log-likelihood, using several artificial and real datasets.**

I. Introduction

Statistical Machine Learning (SML) provides a unified principled framework for machine learning methods and helps to overcome some of their limitations. Bayesian probability theory, in particular, has important modeling implications. For instance, it requires modeling assumptions, including the specification of prior distributions, to be made explicit, avoiding arbitrary modelling decisions; it also automatically satisfies the likelihood principle and provides a natural framework to handle uncertainty.

Generative Topographic Mapping (GTM) [1] is a SML manifold learning model for data visualization and clustering, whose probabilistic setting and functional similarity make it a principled alternative to Self-Organizing Maps (SOM) [2]. In its basic formulation, the GTM is trained within the ML framework using EM, permitting the occurrence of data overfitting unless regularization is included, a major drawback when modelling noisy data. Its probabilistic definition, though, allows the formulation of principled extensions, such as those providing active model regularization. Some regularization methods for GTM described in [3], [4] are based on Bayesian evidence approaches. Alternatively, a variational Bayesian approach of the GTM was recently introduced in [5], [6] to endow the model with regularization capabilities based on variational techniques.

In this paper the performance of Variational GTM is assessed in several experiments, using both artificial and real datasets. Such performance is also compared, in terms of generalization capability (i.e., the capability to avoid overfitting), to that of other GTM models including alternative evidence-based regularization methods, as well as to that of the standard unregularized GTM and the GTM with GP prior.

The remaining of the paper is organized as follows: First, in section II, an introduction to the original GTM, the GTM regularized models based on evidence, the GTM with GP prior and a Bayesian approach for the GTM, are provided. This is followed, in section III, by the description of the Variational GTM. Several experiments for the assessment of the performance of the models are described, and their results presented and discussed, in section IV. The paper wraps up with a brief conclusion section.

Iván Olier and Alfredo Vellido are with the Department of Computing Languages and Systems, Technical University of Catalonia, C/. Jordi Girona 1-3, Edifici Omega, 08034 - Barcelona, Spain (email: {iaolier,avellido}@lsi.upc.edu).

II. Generative Topographic Mapping

A. The Original GTM

The neural network-inspired GTM is a nonlinear latent variable model of the manifold learning family, with sound foundations in probability theory. It performs simultaneous clustering and visualization of the observed data through a nonlinear and topology-preserving mapping from a visualization latent space in $\Re^L$(with L being usually 1 or 2 for visualization purposes) onto a manifold embedded in the $\Re^D$ space, where the observed data reside. The mapping that generates the manifold is carried out through a *regression function* given by:

$$\mathbf{y} = \mathbf{W}\mathbf{\Phi}(\mathbf{u}) \tag{1}$$

where $\mathbf{y} \in \Re^D$, $\mathbf{u} \in \Re^L$, $\mathbf{W}$ is the matrix that generates the mapping, and $\mathbf{\Phi}$ is a matrix with the images of S basis functions ϕ_s (defined as radially symmetric Gaussians in the original formulation of the model). To achieve computational tractability, the prior distribution of $\mathbf{u}$ in latent space is constrained to form a uniform discrete grid of K centres, analogous to the layout of the SOM units, in the form:

$$p(\mathbf{u}) = \frac{1}{K}\sum_{k=1}^{K}\delta(\mathbf{u} - \mathbf{u}_k)$$

This way defined, the GTM can also be understood as a constrained mixture of Gaussians. A density model in data space is therefore generated for each component k of the mixture, which, assuming that the observed data set $\mathbf{X}$ is constituted by $\mathbf{N}$ independent, identically distributed (i.i.d.) data points $\mathbf{x}_n$, leads to the definition of a complete likelihood in the form:

978-1-4244-1821-3/08/$25.00©2008 IEEE

$$P(\mathbf{X}|\mathbf{W},\beta) = \left(\frac{\beta}{2\pi}\right)^{ND/2} \prod_{n=1}^{N} \left\{ \frac{1}{K} \sum_{k=1}^{K} \exp\left(-\frac{\beta}{2} \|\mathbf{x}_n - \mathbf{y}_k\|^2\right) \right\} \quad (2)$$

where $\mathbf{y}_k = \mathbf{W}\mathbf{\Phi}(\mathbf{u}_k)$ are the reference vectors. From Eq. 2, the adaptive parameters of the model, which are $\mathbf{W}$ and the common inverse variance of the Gaussian components, β, can be optimized by ML using the EM algorithm. Details can be found in [1].

B. GTM Regularized Models

The optimization of Eq. 2 makes the model fit whatever noise is present in the dataset. An advantage of the probabilistic definition of the GTM is the possibility of introducing regularization in the mapping. This procedure automatically regulates the level of map smoothing necessary to avoid data overfitting, resorting to either a *single regularization term* [3], or to multiple ones (in a procedure called *Selective Map Smoothing* : [4]). The first case entails the definition of a *penalized* log-likelihood of the form:

$$\ell_{\text{PEN}}(\mathbf{W},\beta) = \ell(\mathbf{W},\beta) - \frac{1}{2}\gamma \|\mathbf{w}\|^2$$

where $\ell(\mathbf{W},\beta)$ is the log-likelihood of the original formulation of GTM (logarithm of Eq. 2), γ is a regularization coefficient and $\mathbf{w}$ is a vector shaped by concatenation of the different column vectors of the weight matrix $\mathbf{W}$.

A Bayesian approach to the estimation of the regularization coefficient γ, as well as the inverse variance β, was introduced in [7]. In this procedure, Bayes' theorem is used to estimate the distribution of γ and β given the data points:

$$p(\gamma,\beta|\mathbf{X}) = \frac{p(\mathbf{X}|\gamma,\beta)\, p(\gamma,\beta)}{p(\mathbf{X})} \quad (3)$$

Assuming uninformative priors, the optimization of the equation 3 is equivalent to the maximization of the *evidence* or marginal likelihood:

$$p(\mathbf{X}|\gamma,\beta) = \int p(\mathbf{X}|\mathbf{w},\beta)\, p(\mathbf{w}|\gamma)\, d\mathbf{w} \quad (4)$$

A normal prior is choosen for the weights:

$$p(\mathbf{w},\gamma) = \left(\frac{\gamma}{2\pi}\right)^{W/2} \exp\left(-\frac{1}{2}\gamma \|\mathbf{w}\|^2\right)$$

where W is the number of weights in $\mathbf{W}$. The log-evidence or marginal log-likelihood for γ and β is given by:

$$\ln p(\mathbf{X}|\gamma,\beta) = \ell(\mathbf{W}_*,\beta) - \frac{1}{2}\gamma \|\mathbf{w}_*\|^2 - \frac{1}{2}\ln|\mathbf{H}_*| + \frac{W}{2}\ln\gamma + C \quad (5)$$

where $\mathbf{W}_*$ is the value of $\mathbf{w}$ at the maximum of the posterior distribution (Eq. 4) and $\mathbf{H}_*$ is the Hessian of $p(\mathbf{X}|\mathbf{w}_*,\beta)\, p(\mathbf{w}_*|\gamma)$. All the constant terms have been grouped as C. The maximization of this equation for γ and β leads to the standard updating formulae of the evidence approximation.

Alternatively, multiple regularization terms can also be considered, one for each basis function. This method known as *Selective Map Smoothing* (SMS) was originally introduced in [4]. In SMS, the prior distribution over the weights is given by

$$p(\mathbf{w},\{\gamma_s\}) = \prod_{s=1}^{S} \left(\frac{\gamma_s}{2\pi}\right)^{D/2} \exp\left(-\frac{1}{2}\sum_{s=1}^{S}\gamma_s \|\mathbf{w}_s\|^2\right)$$

where each γ_s defines a regularization coefficient for each basis function, and $\mathbf{w}_s$ is the vector of weights in $\mathbf{W}$ associated with the hyperparameter s. The marginal log-likelihood of Eq. 5 is reformulated as:

$$\ln p(\mathbf{X}|\{\gamma_s\},\beta) = \ell(\mathbf{W}_*,\beta) - \frac{1}{2}\sum_{s=1}^{S}\gamma_s \|\mathbf{w}_{*s}\|^2 - \frac{1}{2}\ln|\mathbf{H}_*\{\gamma_s\}| + \frac{D}{2}\sum_{s=1}^{S}\ln\gamma_s$$

C. A Gaussian Process Formulation of GTM

The original formulation of GTM described in the previous section has a hard constraint imposed on the mapping from the latent space to the data space due to the finite number of basis functions used. An alternative approach is introduced in [3], where the regression function using basis functions is replaced by a smooth mapping carried out by a GP prior. This way, the likelihood takes the form:

$$P(\mathbf{X}|\mathbf{Z},\mathbf{Y},\beta) = \left(\frac{\beta}{2\pi}\right)^{ND/2} \prod_{n=1}^{N}\prod_{k=1}^{K} \left\{\exp\left(-\frac{\beta}{2}\|\mathbf{x}_n - \mathbf{y}_k\|^2\right)\right\}^{z_{kn}} \quad (6)$$

where: $\mathbf{Z} = \{z_{kn}\}$ are binary membership variables complying with the restriction $\sum_{k=1}^{K} z_{kn} = 1$ and $\mathbf{y}_k = (y_{k1},\ldots,y_{kD})^T$ are the column vectors of a matrix $\mathbf{Y}$ and the centroids of spherical Gaussian generators equivalent to the reference vectors in the case of the orginal formulation of GTM. Note that the spirit of $\mathbf{y}_k$ in this approach is similar to the regression version of GTM (Eq. 1) but with a different formulation: A GP formulation is assumed introducing a prior multivariate Gaussian distribution over $\mathbf{Y}$ defined as:

$$P(\mathbf{Y}) = (2\pi)^{-KD/2}\,|\mathbf{C}|^{-D/2} \prod_{d=1}^{D} \exp\left(-\frac{1}{2}\mathbf{y}_{(d)}^T\mathbf{C}^{-1}\mathbf{y}_{(d)}\right)$$

where $\mathbf{y}_{(d)}$ is each one of the row vectors of the matrix $\mathbf{Y}$ and $\mathbf{C}$ is a matrix where each of its elements is a covariance function that can be defined as

$$\mathbf{C}(i,j) = \mathbf{C}(\mathbf{u}_i, \mathbf{u}_j) = \nu \exp\left(-\frac{\|\mathbf{u}_i - \mathbf{u}_j\|^2}{2\alpha^2}\right), \quad i,j = 1 \dots K$$

and where parameter ν is usually set to 1. The α parameter controls the flexibility of the mapping from the latent space to the data space. An extended review of covariance functions can be found in [8]. An alternative GP formulation was introduced in [9], but this approach had the disadvantage of not preserving the topographic ordering in latent space, being therefore inappropiate for data visualization purposes.

Note that Eqs. 2 and 6 are equivalent if a prior multinomial distribution over $\mathbf{Z}$ in the form $P(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\left(\frac{1}{K}\right)^{z_{kn}} = \frac{1}{K^N}$ is assumed.

Eq. 6 leads to the definition of a log-likelihood, and parameters $\mathbf{Y}$ and β of this model can be optimized using the EM algorithm (in a similar way to the parameters $\mathbf{W}$ and β in the regression formulation of GTM). Some basic details are provided in [3].

D. Bayesian GTM

The specification of a full Bayesian model of GTM can be completed by defining priors over the parameters $\mathbf{Z}$ and β. Since z_{kn} are defined as binary values, a multinomial distribution can be chosen for $\mathbf{Z}$:

$$P(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K} p_{kn}^{z_{kn}}$$

where p_{kn} is the parameter of the distribution.

As in [10], a Gamma distribution[1] is chosen to be the prior over β:

$$P(\beta) = \Gamma(\beta | d_\beta, s_\beta)$$

where d_β and s_β are the parameters of the distribution. Therefore, the joint probability $P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}, \beta)$ is given by:

$$P(\mathbf{X}, \mathbf{Z}, \mathbf{Y}, \beta) = P(\mathbf{X}|\mathbf{Z}, \mathbf{Y}, \beta)\, P(\mathbf{Z})\, P(\mathbf{Y})\, P(\beta)$$

In general, the joint probability can be maximized through evidence methods using the Laplace approximation [7] or, alternatively, using approximate methods, such as Markov Chain Monte Carlo [11] and variational inference [12], [13].

The latter is the approach we follow to define Variational GTM in section III.

III. Variational GTM

A. Motivation of the Use of Variational Inference

A basic problem in SML is the computation of the marginal likelihood $P(\mathbf{X}) = \int P(\mathbf{X}, \Theta)\, d\Theta$, where $\Theta = \{\theta_i\}$ is the set of parameters defining the model. Depending of the complexity of the model, the analytical computation of this integral could be intractable. Variational inference allows approximating the marginal likelihood through Jensen's inequality as follows:

$$\begin{aligned}\ln P(\mathbf{X}) &= \ln \int P(\mathbf{X}, \Theta)\, d\Theta \\ &= \ln \int Q(\Theta) \frac{P(\mathbf{X}, \Theta)}{Q(\Theta)} d\Theta \\ &\geq \int Q(\Theta) \ln \frac{P(\mathbf{X}, \Theta)}{Q(\Theta)} d\Theta = F(Q)\end{aligned}$$

The function $F(Q)$ is a lower bound function such that its convergence guarantees the convergence of the marginal likelihood. The goal in variational methods is choosing a suitable form for the density $Q(\Theta)$ in such a way that $F(Q)$ can be readily evaluated and yet which is sufficiently flexible that the bound is reasonably tight. A reasonable approximation for $Q(\Theta)$ is based on the assumption that it factorizes over each one of the parameters as $Q(\Theta) = \prod_i Q_i(\theta_i)$. That assumed, $F(Q)$ can be maximized leading the optimal distributions:

$$Q_i(\theta_i) = \frac{\exp\langle \ln P(\mathbf{X}, \Theta)\rangle_{k \neq i}}{\int \exp\langle \ln P(\mathbf{X}, \Theta)\rangle_{k \neq i}\, d\theta_i} \tag{7}$$

where $\langle\, .\, \rangle_{k \neq i}$ denotes an expectation with respect to the distributions $Q_k(\theta_k)$ for all $k \neq i$.

B. A Bayesian Approach of GTM Based on Variational Inference

In order to apply the variational principles to the Bayesian GTM within the framework described in the previous section, a Q distribution of the form:

$$Q(\mathbf{Z}, \mathbf{Y}, \beta) = Q(\mathbf{Z})\, Q(\mathbf{Y})\, Q(\beta)$$

is assumed, where natural choices of $Q(\mathbf{Z})$, $Q(\mathbf{Y})$ and $Q(\beta)$ are similar distributions to the priors $P(\mathbf{Z})$, $P(\mathbf{Y})$ and $P(\beta)$, respectively. Thus, $Q(\mathbf{Z}) = \prod_{n=1}^{N}\prod_{k=1}^{K}\tilde{p}_{kn}^{z_{kn}}$, $Q(\mathbf{Y}) = \prod_{d=1}^{D}\mathcal{N}\left(\mathbf{y}_{(d)}|\tilde{\mathbf{m}}^{(d)}, \tilde{\mathbf{\Sigma}}\right)$, and $Q(\beta) = \Gamma\left(\beta|\tilde{d}_\beta, \tilde{s}_\beta\right)$. Using these expressions in Eq. 7, the following formulation for the variational parameters $\tilde{\mathbf{\Sigma}}, \tilde{\mathbf{m}}^{(d)}, \tilde{p}_{kn}, \tilde{d}_\beta$ and $\tilde{s}_\beta$ can be obtained:

[1] The Gamma distribution is defined as follows: $\Gamma(\nu|d_\nu, s_\nu) = \frac{s_\nu^{d_\nu}\nu^{d_\nu - 1}\exp^{-s_\nu \nu}}{\Gamma(d_\nu)}$

$$
\begin{aligned}
\tilde{\boldsymbol{\Sigma}} &= \left(\langle \beta \rangle \sum_{n=1}^{N} \mathbf{G}_n + \mathbf{C}^{-1} \right)^{-1} \\
\tilde{\mathbf{m}}^{(d)} &= \langle \beta \rangle \, \tilde{\boldsymbol{\Sigma}} \sum_{n=1}^{N} x_{nd} \, \langle \mathbf{z}_n \rangle \\
\tilde{p}_{kn} &= \frac{\exp\left\{ -\frac{\langle \beta \rangle}{2} \left\langle \left\| \mathbf{x}_n - \mathbf{y}_k \right\|^2 \right\rangle \right\}}{\sum_{k'=1}^{K} \exp\left\{ -\frac{\langle \beta \rangle}{2} \left\langle \left\| \mathbf{x}_n - \mathbf{y}_{k'} \right\|^2 \right\rangle \right\}} \\
\tilde{d}_\beta &= d_\beta + \frac{ND}{2} \\
\tilde{s}_\beta &= s_\beta + \frac{1}{2} \sum_{n=1}^{N} \sum_{k=1}^{K} \langle z_{kn} \rangle \left\langle \left\| \mathbf{x}_n - \mathbf{y}_k \right\|^2 \right\rangle
\end{aligned}
$$

where $\mathbf{z}_n$ corresponds to each row vector of $\mathbf{Z}$ and $\mathbf{G}_n$ is a diagonal matrix of size $K \times K$ with elements $\langle \mathbf{z}_n \rangle$. The moments in the previous equations are defined as: $\langle z_{kn} \rangle = \tilde{p}_{kn}$, $\langle \beta \rangle = \frac{\tilde{d}_\beta}{\tilde{s}_\beta}$, and $\left\langle \left\| \mathbf{x}_n - \mathbf{y}_k \right\|^2 \right\rangle = D\tilde{\boldsymbol{\Sigma}}_{kk} + \sum_{d=1}^{D} \left(x_{nd} - \tilde{m}^{(kd)} \right)^2$.

Details of these calculations can be found in [5].

IV. Experiments

A. Experimental Design

The main goal of the set of experiments presented and discussed in this section is the assessment of the performance of the proposed Variational GTM in the presence of noise. That is, the assessment of its robustness in terms of model regularization. The performance of the Variational GTM is compared with those of the original unregularized GTM; the GTM regularized using evidence methods, either with a single regularization term, or with multiple ones; and the GP formulation for GTM.

The models used in all the experiments were initialized in the same way to allow straightforward comparison. The matrix centroids of the Gaussian generators $\mathbf{Y}$ and the inverse of the variance β were set through PCA-based initialization [1] and the parameters $\{p_{kn}\}$ are fixed and were initialized to $1/K$. The parameter s_β was set to d_β/β and d_β was initialized to a small value close to 0. For each set of experiments, several values of α were tried though finally it was set to 0.1.

Five publicly available datasets and a sixth synthetically generated one, all with different characteristics, were selected for the experiments. They are now summarily described:

- *Wine_data*: This dataset consists of 13 attributes and 179 cases, describing the results of chemical analysis of wine samples. It is available from the UCI machine learning repository[2].
- *3-PhaseOil_data*: This dataset consisting of 12 attributes and 1,000 data points was artificially generated from the dynamical equations of a pipeline section carrying a mixture of oil, water and gas which can belong to one of three equally distributed geometrical configurations. It was originally used in [1] and it is available in the GTM Homepage[3].
- *Shuttle_data*: It is a dataset consisting of 6 attributes and 1,000 data points obtained from various inertial sensors from Space Shuttle mission STS-57[4].
- *Abalone_data*: Another dataset from the UCI repository consisting of 8 attributes and 3,175 data points. It was originally used to predict the age of abalone marine gastropods from physical measurements.
- *Letter_data*: This dataset consists of 16 attributes and 20,000 data points, used for letter category recognition. It is also available from the UCI repository.
- *Spiral_data*: A simple two-dimensional artificial dataset consisting of 200 data points was artificially generated using the equation of a spiral contaminated with Gaussian noise, as follows:

$$
\mathbf{X} = \left[\begin{array}{l} x_1 = \frac{n}{200} \sin\left(4\pi n/200\right) + \sigma\left(0.05\right) \\ x_2 = \frac{n}{200} \cos\left(4\pi n/200\right) + \sigma\left(0.05\right) \end{array} \right],
$$

where $1 \leq n \leq 200$ and $\sigma\left(0.05\right)$ is the Gaussian noise with standard deviation of 0.05.

B. Comparative Assessment of the performance of Variational GTM

The performance of all methods is assessed using the test log-likelihood of the resulting models. Ten-fold cross-validation for each dataset and method was used. The results of the experiments are shown in Figs. 1 to 6. These figures summarily display the test log-likelihoods for each method, as a function of the number of latent points. All figures provide evidence that the proposed Variational GTM outperforms the rest of models, overall (with the exception of the *Shuttle_data*) and for almost any number of latent points. Moreover, this difference of performance is, in some cases (Figs. 1, 3, 5 and 6), quite big. In contrast with other models (such as GTM-GP in Figs. 1 and 2), the performance of Variational GTM does not deteriorate with the number of latent points. Interestingly, the performance of the original GTM and the GTM regularized with evidence-based methods (GTM-SRT and GTM-SMS) is quite similar in all figures. In turn, in most cases, the performances of evidence-based methods and GTM-GP are very similar up to a number of latent points, beyond which their performances diverge notably.

C. On the influence of Model Regularization in the Visualization of the Data

The low dimensionality of the *Spiral_data* set allows us to display it directly in Fig. 7, together with the corresponding reference vectors $\mathbf{y}_k$ obtained using each of the GTM variants. The original spiral without noise is also added to the displays so that the level of fitting of each model to the data can be visually assessed. It is clearly observed

[2] http://mlearn.ics.uci.edu/MLRepository.html

[3] http://www.ncrg.aston.ac.uk/GTM

[4] http://www.cs.ucr.edu/∼eamonn/

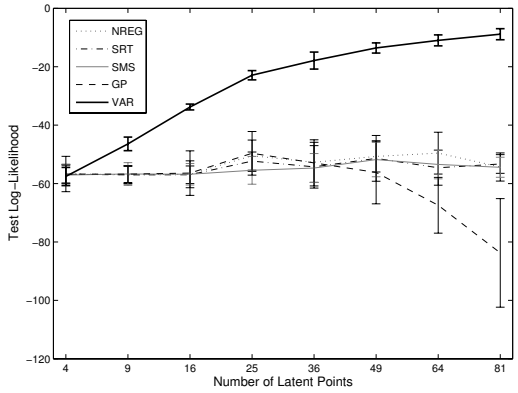

Fig. 1. Mean test log-likehood results for the *Spiral_data* for all methods: Unregularized GTM (NREG); GTM regularized with evidence methods: Single regularization term (SRT) and Selective Mapping Smoothing (SMS); GTM with GP prior (GP); and Variational GTM (VAR). The vertical bars indicate the standard deviation of the test log-likelihood over the cross-validation runs.

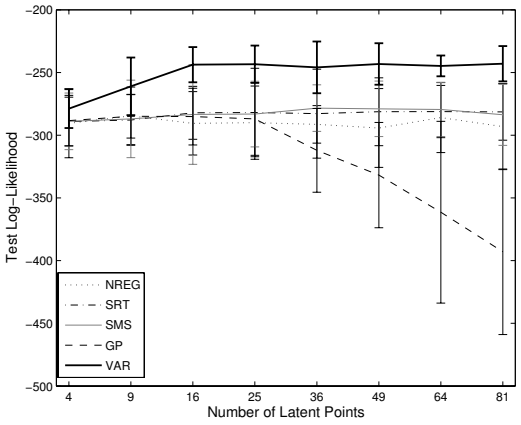

Fig. 2. Mean test log-likehood results for the *Wine_data*. Representation as in Fig. 1.

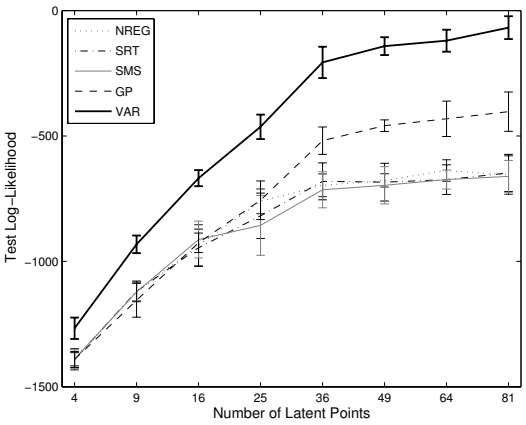

Fig. 3. Mean test log-likehood results for the *3-PhaseOil_data*. Representation as in Fig. 1.

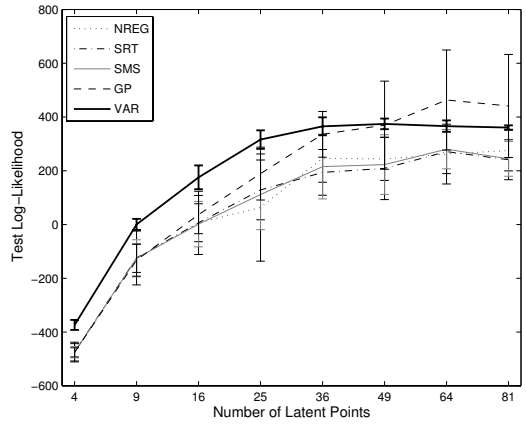

Fig. 4. Mean test log-likehood results for the *Shuttle_data*. Representation as in Fig. 1.

that the Variational GTM approximates the original spiral far better than any of the alternative methods (leading to better generalization capabilities, as illustrated by the test log-likelihood results reported in the previous section), which tend to be more sensible to the effect of the added noise (allocating, as a result, some reference vectors to areas outside the original spiral).

For data of higher dimensionality, two visualization strategies can be followed. In the first one, data are visualized in two dimensions in the model latent space, using the mean projection [1] calculated as $\mathbf{u}_n^{\texttt{mean}} = \sum_k p(\mathbf{u}_k|\mathbf{x}_n)\,\mathbf{u}_k$ for all methods with exception to the Variational GTM, for which is calculated as $\mathbf{u}_n^{\texttt{mean}} = \sum_k \langle z_{kn}\rangle \mathbf{u}_k$. This is illustrated by the visualization of the *Wine_data* set. The original dataset was first divided into a training subset (66% of all data points, randomly selected) and a test subset (rest of the data). The training data are visualized for all GTM variants in Fig. 8, while the test data are visualized in Fig. 9. Both figures show that, for all models but Variational GTM, the data ocupy most of the latent space. Thus, their visualization does not reveal any clear grouping structure. The original three-class structure of the *Wine_data* is only recognized by labelling each class differently in the display. Instead, Variational GTM captures the underlying three-class structure perfectly, isolating each group in a very defined area of the latent space. Moreover, the labelling of data points allows us to identify, without any ambiguity, several data points which are clearly mislabeled: that is, points with a class label that does not correspond to their natural grouping as revealed by Variational GTM.

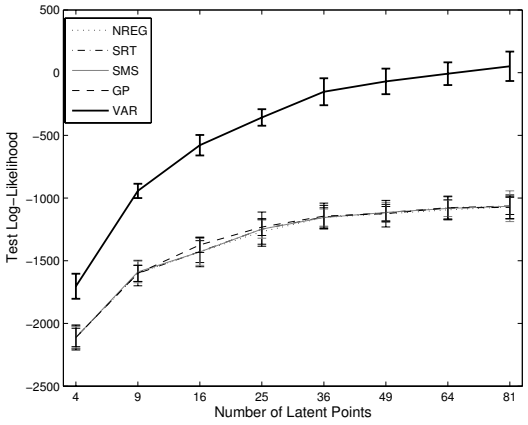

Fig. 5. Mean test log-likehood results for the *Abalone_data*. Representation as in Fig. 1.

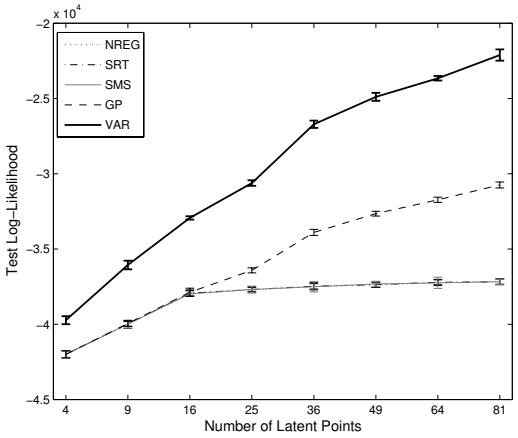

Fig. 6. Mean test log-likehood results for the *Letter_data*. Representation as in Fig. 1.

The second strategy deals with the visualization of the general cluster structure defined by the GTM variants. It is accomplished through the membership map generated using the mode projection [1] of the data into the latent space, given by $\mathbf{u}_n^{\mathtt{mode}} = \underset{k}{\mathtt{argmax}}\ p(\mathbf{u}_k|\mathbf{x}_n)$ for all methods with exception to the Variational GTM, for which is given by $\mathbf{u}_n^{\mathtt{mode}} = \underset{k}{\mathtt{argmax}}\ \langle z_{kn} \rangle$. This is illustrated by the visualization of the *Wine_data* set clusters in Figs. 10 and 11. Again, as in the case of the mean projections, the underlying three-class structure of the data is only clearly observed in Variational GTM. Moreover, only Variational GTM provides a parsimonious cluster description of the data, using a very small number of clusters for each of the three wine classes. This reflects the success of the regularization process. In comparison, the rest of GTM variants, regularized or not,

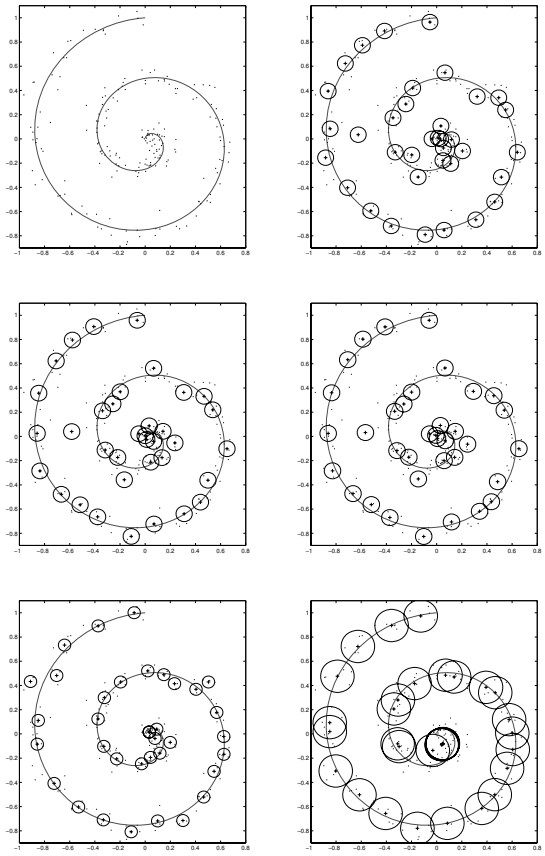

Fig. 7. (Top row, left) *Spiral_data*, (Top row, right) original GTM, (Middle row, left) GTM-SRT, (Middle row, right) GTM-SMS, (Bottom row, left) GTM-GP, and (Bottom row, right) Variational GTM. The common standard deviation is represented by circles centred on each reference vector, with radius $1/\sqrt{\beta}$.

show a proliferation of clusters that is the result of data overfitting.

V. Conclusions

The benefits of a Variational formulation for the manifold learning GTM model, in order to achieve effective model regularization, have been demostrated in this paper. Several experiments, using diverse datasets of very different characteristics, have shown that Variational GTM is able to avoid, at least partially, data overfitting and, therefore, is able to generalize better than several alternative GTM formulations, both regularized and unregularized. Additionaly, the advantages of the variational formulation for data and cluster visualization have been clearly illustrated.

Future research will be devoted to include some other model parameters within the variational framework. In particular, a variational treatment of hyperparameter α is difficult. However, an interesting approach to its calculation in the context of variational GP classifiers, using lower and upper

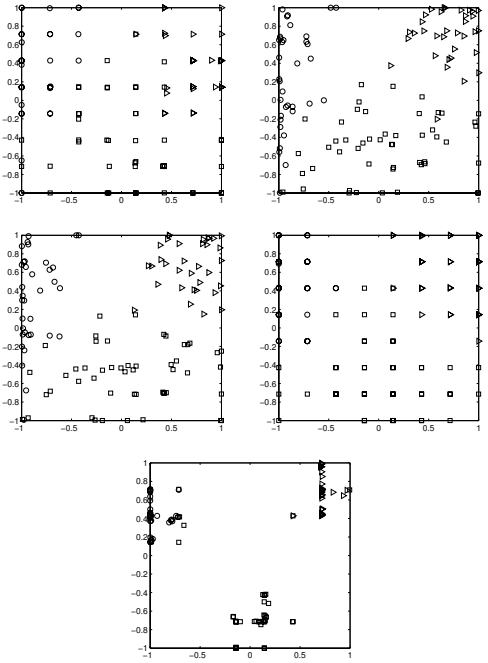

Fig. 8. Data visualization through mean projection for the training subset of the *Wine_data*, (Top row, left) original GTM, (Top row, right) GTM-SRT, (Middle row, left) GTM-SMS, (Middle row, right) GTM-GP, and (Bottom row) Variational GTM.

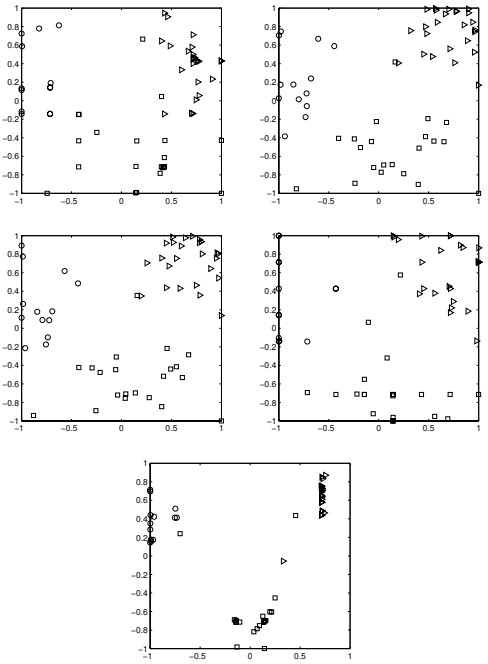

Fig. 9. Data visualization through mean projection for the test subset of the *Wine_data*, (Top row, left) original GTM, (Top row, right) GTM-SRT, (Middle row, left) GTM-SMS, (Middle row, right) GTM-GP, and (Bottom row) Variational GTM.

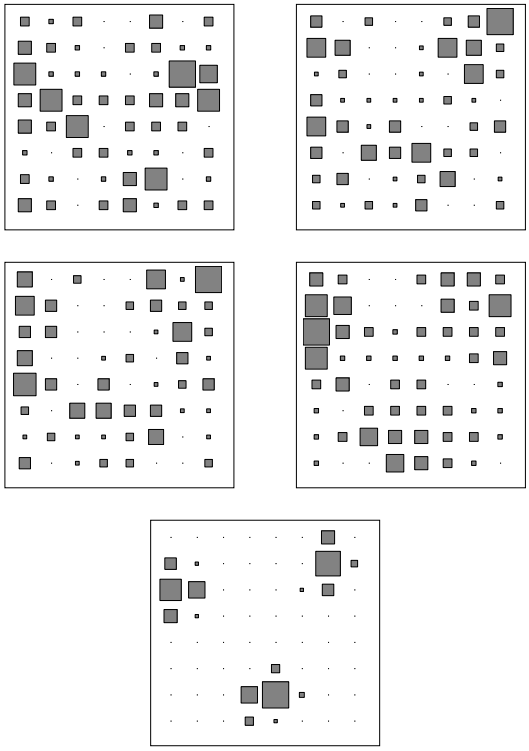

Fig. 10. Data visualization through membership maps for the training subset of the *Wine_data*, (Top row, left) original GTM, (Top row, right) GTM-SRT, (Middle row, left) GTM-SMS, (Middle row, right) GTM-GP, and (Bottom row) Variational GTM. Each cluster is represented by a square of size proportional to the number of data points assigned to it.

bound funtions, was presented in [14] and will be explored in the context of GTM. Furthermore, an additional vector of adaptive hyperparameters over parameter $\mathbf{Y}$ could be used to control the mixture of Gaussian components. Thereby, an optimum number of mixture components could be calculated.

Finally, we remark that the computational complexity of Variational GTM does not increase with respect to that of the standard GTM with GP prior. On the other hand, the formulation of Variational GTM introduces a heavier computational load as compared to the standard GTM, as usual in most formulations involving Bayesian inference. However, there was no significant increase in the running times for the experiments reported in this paper. A more thorough study of the computational efficiency of the method will also be the matter of future research.

REFERENCES

[1] C. M. Bishop, M. Svensén, and C. K. I. Williams, "GTM: The Generative Topographic Mapping," *Neural Comput.*, vol. 10, no. 1, pp. 215–234, 1998.

[2] T. Kohonen, *Self-Organizing Maps (3rd ed)*. Springer-Verlag, Berlin, 2001.

[3] C. M. Bishop, M. Svensén, and C. K. I. Williams, "Developments of the Generative Topographic Mapping," *Neurocomputing*, vol. 21, no. 1–3, pp. 203–224, 1998.

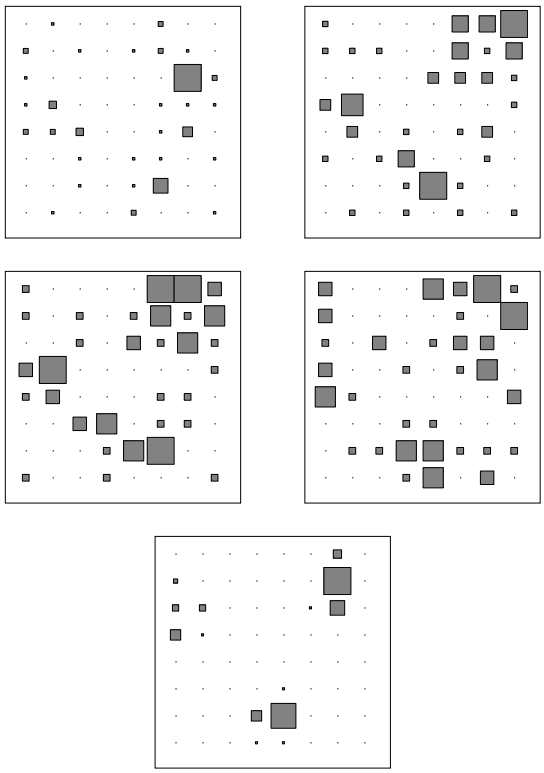

Fig. 11. Data visualization through membership maps for the test subset of the *Wine_data*, (Top row, left) original GTM, (Top row, right) GTM-SRT, (Middle row, left) GTM-SMS, (Middle row, right) GTM-GP, and (Bottom row) Variational GTM. Cluster representation as in Fig. 10.

[4] A. Vellido, W. El-Deredy, and P. J. G. Lisboa, "Selective smoothing of the Generative Topographic Mapping," *IEEE T. Neural Networ.*, vol. 14, no. 4, pp. 847–852, 2003.

[5] I. Olier and A. Vellido, "A variational Bayesian formulation for GTM: Theoretical foundations," Technical University of Catalonia (UPC), Tech. Rep. LSI-07-33-R, 2007.

[6] ——, "Variational GTM," in *The 8th International Conference on Intelligent Data Engineering and Automated Learning (IDEAL'07). Lect. Notes Comput. Sc.*, vol. 4881, 2007, pp. 77–86.

[7] D. J. C. MacKay, "A practical Bayesian framework for back-propagation networks," *Neural Comput.*, vol. 4, no. 3, pp. 448–472, 1992.

[8] P. Abrahamsen, "A review of Gaussian random fields and correlation functions," Norwegian Computing Center, Oslo, Norway, Tech. Rep. 917, 1997.

[9] A. Utsugi, "Bayesian sampling and ensemble learning in Generative Topographic Mapping," *Neural Process. Lett.*, vol. 12, pp. 277–290, 2000.

[10] C. M. Bishop, "Variational principal components," in *Proceedings Ninth Intern. Conf. on Artificial Neural Networks*, vol. 1, 1999, pp. 509–514.

[11] C. Andrieu, N. de Freitas, A. Doucet, and M. I. Jordan, "An introduction to MCMC for machine learning," *Mach. Learn.*, vol. 50, pp. 5–43, 2003.

[12] M. Beal, "Variational algorithms for approximate Bayesian inference," Ph.D. dissertation, The Gatsby Computational Neuroscience Unit, Univ. College London, 2003.

[13] T. Jakkola and M. I. Jordan, "Bayesian parameter estimation via variational methods," *Stat. Comput.*, vol. 10, pp. 25–33, 2000.

[14] M. Gibbs and D. J. C. MacKay, "Variational Gaussian process classifiers," *IEEE T. Neural Networ.*, vol. 11, no. 6, pp. 1458–1464, 2000.